Understandable and Singable Musical Lyrics Translation

Anonymous ACL submission

Abstract

Translating lyrics in musicals is a new and challenging task due to various constraints to consider. While previous song translation works explore ways to incorporate and satisfy music constraints, they cannot ensure basic translation quality which is critical in musicals. This paper is dedicated to enhancing translation quality while simultaneously maintaining the singability features, such as adherence to length and rhyme constraints. Our approach consists of three main components. First, we collect a dataset to train reward models, giving an automatic evaluation of translation quality. To enhance both the singability and translation capabilities, we adopt a two-stage training recipe with filtering techniques. Finally, our inference-time optimizing framework composes the whole-song translation. Extensive experiments of both automatic evaluations and human evaluation not only show improvements over the baseline method by large margins but also demonstrate the effectiveness of different components in our approach.

1 Introduction

Have you ever heard of Hamilton in Chinese[1], or Mamma Mia in Swedish (Åkerström, 2010)? Advancements in cultural globalization allow musicals to cross national borders and reach audiences worldwide, but language barriers still pose a challenge for audiences to fully understand the show. Translating musicals to the performing country's language is a good way to enhance audiences' experience (Sorby et al., 2014) and win a bigger market (Andersson et al., 2008), as it enables audiences to enjoy rich theatre components in musicals (Engel and Kissel, 2006) instead of heavily focusing on subtitles (Sorby et al., 2014).

However, musical translation is labor-intensive and time-consuming because translators need to not only translate text but also consider its compatibility with music framework, stage performance, and cultural references (Sorby et al., 2014; Fei, 2014). In the hope of alleviating their burden, we aim to automatically translate lyrics in musical songs from one language to another. We consider translation from English to Chinese in our work, but our method can be generalized to other languages as well.

Regarding song translation, several well-recognized translatology theories discussed the criteria to consider (Low, 2003; Franzon, 2005), which can be summarized as text-music alignment and translation quality. To the best of our knowledge, existing arts on neural song translation (Guo et al., 2022; Ou et al., 2023; Li et al., 2023a) focus on the alignment of text and music. Although they achieved decent results in making translated lyrics comfortable for singing, they sacrifice translation quality and often produce unnatural and inaccurate translations, as shown in the baseline results of Figure 1. These ill-quality lyrics are especially unsuitable for musicals since musical lyrics are all functional and serve as part of the "storytelling" role (Kenrick, 2010; Carpi, 2020; Chan, 2017). Thus, the lyrics must be understandable and of high literal quality.

In comparison, our work focuses on improving translation quality while following the singability constraints. We define translation quality with the well-known criteria for translation: fluency, accuracy, and literacy (Yan, 1898). We consider the singable constraints of length and rhyme following previous works (Guo et al., 2022; Ou et al., 2023). Figure 1 shows our considered aspects, with examples demonstrating their significance.

To depict translation quality, we collect English-Chinese lyrics pairs using large language models (LLMs), label them according to our scoring

[1]The Chinese version of Hamilton translated by musical enthusiasts: https://www.bilibili.com/video/BV1YN411873B/

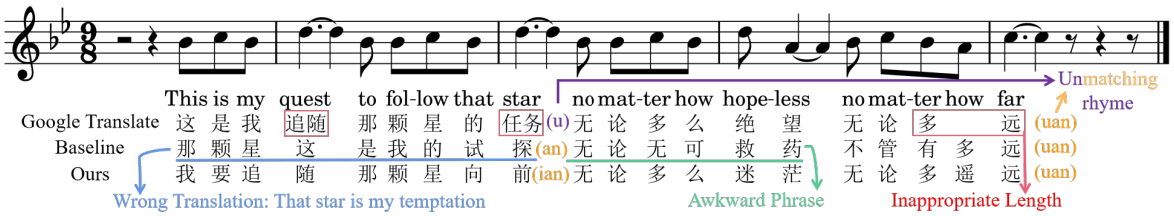

Figure 1: Constraints we considered: length, rhyme, and translation quality. The proper length of translated lyrics is the number of notes, and the end rhyme of each line (shown in parentheses) is better to have the same type (shown in the same color). Google translation fails to follow the length constraint and misaligns with music, as shown in red boxes, and its rhyme does not match either. Both baseline and our results meet length and rhyme constraints, but the baseline has inaccurate translations and inappropriate phrases, while our model generates higher-quality lyrics.

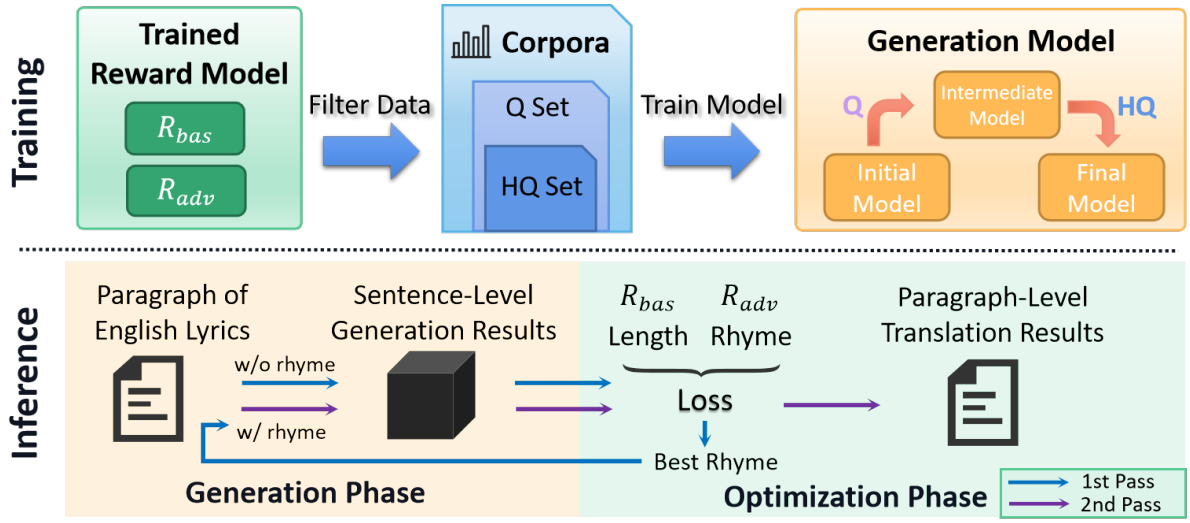

Figure 2: Overview of our pipeline. There are three key components in our method: reward model for subjective aspects evaluations, generation model fine-tuning, and inference-time optimization framework. We use reward models to filter the whole corpora into a **Q**uality subset and a **H**igh-**Q**uality subset and train our LLM first with Q set and then with HQ set. During inference, we generate a large amount of sentence-level translations and derive the paragraph-level translations by optimizing the loss function considering various aspects. To better ensure the same rhyme usage in a paragraph, we use a 2nd pass to perform the same process again with rhyme conditioning.

rubrics, and train reward models to give evaluations that correlate with human scoring. For the singable constraints, we find that LLMs cannot follow the constraints in a zero-shot way so we tune our model to improve the accuracy, while at the same time achieving a balance with translation quality using filtered high-quality data and multi-phase training. Finally, to compose decent translations for the whole passage, we propose an inference-time optimization framework, utilizing our evaluation of paragraph translations and the stochasticity of an LLM. Extensive experiments and analyses have shown the effectiveness of the various components of our method, significantly outperforming the previous state-of-the-art approach.

2 Related Work

Translatology: Song and Musical Translation. In translatology, "Pentathlon Principle" (Low, 2003, 2005) is a wide-known theory and guidance on general song translation (Franzon, 2008; Cheng, 2013; Stopar, 2016; Si-yang, 2017; Opperman et al., 2018; Sardiña, 2021; Pidhrushna, 2021; Ou et al., 2023), which proposes five criteria to consider: singability, rhyme, rhythm, sense and naturalness, where the first three relates to music-text alignment and the rest refer to translation quality. However, this principle is not developed specifically for songs on musical stage (Carpi, 2020).

The functional approach (Franzon, 2005) is more suitable for songs in musicals (Carpi, 2020), which emphasizes that the translated lyrics should replicate the function of the source text. In musi-

cals, songs are "story-telling" elements (Kenrick, 2010), and the translated lyrics must carry out this role (Desblache, 2018; Åkerström, 2010; Sorby et al., 2014; Franzon, 2005). Thus a basic yet necessary constraint in musical translation is that lyrics must be understandable and of high quality.

Automatic Song Translation. To our best knowledge, there are only three previous works on automatic song translation (Guo et al., 2022; Ou et al., 2023; Li et al., 2023a). Guo et al. (2022) mainly addresses the problem of aligning words' tones with the melody in the beam search phase, and Li et al. (2023a) focuses on better aligning text to musical notes. However, they all neglect the important rhyme constraint (Strangways, 1921), which is more critical in comparison. Ou et al. (2023) considers length, rhymes, and word boundaries, and achieves decent results with prompting and the trick of reverse-order decoding. However, the translation quality is awkward and not ready for singing in musicals. Our method considers the two most important constraints in text-music alignment, length, and rhyme, and focuses on generating high-quality translations.

LLM for Machine Translation. Recent years have witnessed the success of large language models (LLMs), like close-sourced GPT-4 (OpenAI, 2023), Kimichat[2], and open-sourced Llama-2 (Touvron et al., 2023). Recent works (Yang et al., 2023; Zhang et al., 2023; Zeng et al., 2024; Chen et al., 2023; Li et al., 2023b; Zhu et al., 2023) sought to enhance the machine translation capability using smaller LLMs, typically 7B or 13B, yet the improvements are quite limited. One challenge is balancing performance improvements during training without significantly compromising the pre-trained model's knowledge. As one contemporary study (Xu et al., 2024) pointed out, there is diminished necessity for parallel data to fine-tune LLMs. Instead, it is recommended to first train with monolingual data if the LLM does not have too much knowledge of our target language, and then fine-tune with a small amount of high-quality parallel data. Though our constrained translation setting is slightly different, we find high-quality parallel data is still very beneficial.

3 Method

Figure 2 gives an overview of our method. To generate high-quality and singable musical translations, we consider various aspects as detailed in Section 3.1. We first collect data and train reward models to evaluate the hard-to-quantize translation quality, as shown in Section 3.2. To ensure good translation quality while maintaining reasonably high length and rhyme accuracy, we design our two-stage training recipe with dataset filtering, introduced in Section 3.3. To derive translation results for a whole paragraph, we propose the inference-time optimization framework in Section 3.4 that further boosts singability and translation quality.

3.1 Controlling Aspects

Our controlling aspects can be categorized into two types: translation quality and text-music alignment. We define translation quality as fluency, accuracy, and literacy (Yan, 1898), and further categorize them into basic and advanced ones. Basic translation quality includes 1) fluency: whether the translation forms a reasonable sentence, and 2) accuracy: whether the translation is faithful to its original meaning. Advanced translation quality is literacy, which describes whether the translation looks like good lyrics, or whether the language usage is elegant.

As for the text-music alignment, or singability aspects, we mainly consider the length constraint and rhyme constraint, which are the most critical two factors for singing the translation results.

3.2 Utilization of the Reward Model

We collect data and train reward models to make evaluations for translation quality that align with human preferences. In particular, for basic translation quality and advanced translation quality, we train reward models and call their evaluation results R_{bas} and R_{adv}, respectively.

Dataset. Every entry of our collected dataset contains an original English line, a translated Chinese line, a paragraph as context, and three scores ranging from 1 to 4 measuring fluency, accuracy, and literacy respectively. We design the detailed rubrics for scoring in collaboration with an expert in the field of musical translation. The English lines are extracted from musicals of diverse genres, as shown in Figure 3(a), while the corresponding Chinese translations are generated by Kimichat with few-shot prompts. After annotating for about 50 hours, we obtain the dataset with 3938 high-quality entries. Please check more details in Appendix A.

Training the Reward Model. We first apply mappings to tackle categories that rarely appear to ob-

[2] https://www.moonshot.cn/

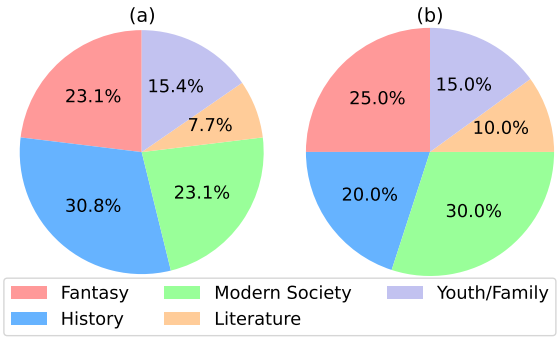

Figure 3: The distribution of the musicals used for annotation (a) and testing (b).

tain a more balanced training dataset for R_{bas} and R_{adv}. For R_{bas}, we map the score pair of fluency and accuracy to a single integer score in the range 1 to 4, obtaining 471, 322, 971, and 2174 entries respectively. For R_{adv}, we map the scores for literacy to 2 or 3 and obtain 3104 and 834 data samples respectively. More details are in Appendix A.

With the data upsampling and downsampling trick to further balance the data in training, we obtained R_{bas} and R_{adv} with strong correlations with humans on the hidden balanced test set that consists of unseen musicals during training time. The Pearson correlation (Pearson, 1895) of humans with R_{bas} and R_{adv} are 0.649 and 0.532, signifying strong and moderate correlation. Besides, the precision and recall of the scoring 3 class R_{adv} are 0.95 and 0.49. The strong correlation of R_{bas} and high precision of R_{adv} make them quite reliable and valuable in our pipeline.

3.3 Two-Stage Training with Dataset Filtering

We observe that a pre-trained LLM cannot follow length and rhyme constraints, and if we want to achieve high length and rhyme accuracy, large-scale training is necessary, as demonstrated by our experiments in Figure 4 (a). However, more training data does not necessarily make translation results better, as demonstrated by both Figure 4 (b), (c). A natural question then arises: how can we ensure good translation quality, with reasonably high length and rhyme accuracy?

Due to the difficulty of collecting a large-scale musical dataset, we adopt the dataset provided by Ou et al. (2023) consisting of about 2.8M song lyrics sentence translations from English to Chinese for training, although there is some gap between normal songs and musical songs. To abridge the gap and improve dataset quality, we use our reward models to filter a high-quality subset of size 1.75M and a higher-quality subset of size 700K. In the first training stage, we train the LLM with the large-scale high-quality dataset mainly to learn length and rhyme constraints. In the second stage, we further refine translation quality by fine-tuning using the higher-quality dataset. In both training stages, we add both the length and rhyme constraints in the prompt. Our later experiments show the effectiveness of our training recipe.

3.4 Inference-Time Optimization Framework

Due to the inaccuracy of generating the whole paragraph all at once, we choose to let the translation model tackle each sentence independently and combine them using a novel optimization framework during inference time. In particular, we design a proper loss function and optimize the overall loss by jointly considering all sentences.

In our setting, we simultaneously consider those aspects: length accuracy, rhyme score, and both basic and advanced translation quality. At the paragraph level, our overall loss $\mathcal{L}(\cdot)$ is defined for sentence-level translations $y_1, \ldots, y_n$ by considering all those aspects. Specifically, we define

$$\mathcal{L}(y_1, \ldots, y_n) = \sum_i (\lambda_1 [\text{Rhy}(y_i) \neq \text{Rhy}(y_n)]$$
$$+\lambda_2 D(\text{gt}_i, |y_i|) - \lambda_3 R_{\text{adv}}(y_i) - \lambda_4 R_{\text{bas}}(y_i))$$

We explain notions here. $D(\cdot, \cdot)$ measures the length difference since we require the translation to have the same length as specified so that we can sing the characters perfectly according to the notes. If the translation is longer than desired, we set a larger penalty than a shorter one since it is more undesired for singing. The two reward models R_{bas} and R_{adv} are introduced earlier, and according to our rubrics, the translation basic quality is a compulsory requirement to ensure acceptable translation results, we only leave those with $R_{\text{bas}} \geq 3$. $\text{Rhy}(\cdot)$ specifies the rhyme type of the last character, and the rhyme grouping rule is from a Chinese music translation book (Xue, 2002).

Then our goal is to find a paragraph translation that approximately minimizes the optimization problem. We choose a proper temperature to call the generation function and draw a lot of translation samples for each sentence so that the generations are diverse and can cover a large portion of high-probability outputs in the generation space. Then the joint choice of $y_1, \ldots, y_n$ has a

Method (Training Config.)	Rhyme	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
Ou et al. (2023)	yes	**0.977**	**0.96**	2.845	2.053	18.01	71.94
Ours VER.1 (1.75M)	yes	0.941	0.722	2.789	2.046	18.22	71.93
	no	0.854	-	2.92	2.053	17.15	71.61
Ours VER.2 (1.75M Q)	yes	0.914	0.687	2.971	2.056	18.32	72.87
	no	0.819	-	3.063	2.059	17.68	72.49
Ours VER.3 (1.75M Q + 700K HQ)	yes	0.923	0.703	**3.168**	**2.063**	18.80	**74.14**
	no	0.874	-	3.248	2.068	17.76	73.78

Table 1: Sentence-level results of the three versions of our method. In VER.1, we train the model with a 1.75M subset. In VER.2, we use a 1.75M **Q**uality subset. In VER.3, we use a 700K **H**igh-**Q**uality subset to fine-tune VER.2 model. Rhyme in the heading row means whether we use the rhyme constraint during inference, and the best results of the two cases are in **bold** (use) and underline (without use), respectively.

Method	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
Ou et al. (2023)	0.962	**0.95**	2.744	2.02	13.81	65.5
Ours VER.1	0.982	0.831	3.627	2.23	13.42	67.54
Ours VER.2	0.992	0.868	3.655	2.248	12.95	67.77
Ours VER.3	**0.99**	0.873	**3.76**	**2.248**	12.32	**69.43**

Table 2: The final whole-song translation results of three versions of our method. Compared with Table 1, our method here also includes the optimization part and can fully demonstrate our strength.

super large number of possibilities. However, due to the nice form of the optimization formula, we can solve it fast by enumerating $\text{Rhy}(y_n)$ first and then optimizing each sentence independently.

After finding the sentences $y_1, \ldots, y_n$ that minimize the loss function, we immediately have a recommended rhyme usage. To ensure as many sentence translations as possible can match the desired rhyme, in the second stage, we generate more samples for each sentence with the rhyme conditioning. This second stage is more focused and more economic in sample efficiency after the desired rhyme is already fixed.

In our framework, the choice of the LLM is flexible and we can boost the overall performance by scaling the number of samples for each sentence. Our framework can also support more fine-grained constraints, which can be useful since there are more minor issues to consider related to singability for the translation results to sound better.

4 Experiments

4.1 Experiment Configurations

Datasets. To evaluate the musical translation performance, we collect our musical datasets, as there is no existing musical dataset available to the knowledge of the authors. Our testing dataset consists of musicals of various genres downloaded from Cloud Music[3], and its distribution is shown in Figure 3 (b). Our dataset, sourced from 56 popular songs in musicals, includes 409 paragraphs and 1741 lines, featuring English lyrics and their Chinese translations. Please check more details of dataset processing in Appendix A.

Models. For both the generation model and the reward model, we choose `Chinese-Alpaca-2-13B`[4] (Cui et al., 2023) as our base model since it is pretrained with a large amount of Chinese corpora with instruction-following dataset.

Baselines. To the best of our knowledge, there are only three previous works on song translation, GagaST (Guo et al., 2022), Controllable Lyric Translation (Ou et al., 2023), and LTAG (Li et al., 2023a). Due to data acquisition difficulties of GagaST and LTAG, we have Ou et al. (2023) as our baseline. We train the baseline model directly using its released code[5].

Metrics. For automatic evaluation, we consider **length accuracy (LA)**, which is defined as the percentage of translated sentences whose length equals the desired length (we set it as the length of reference translation for sentence level testing, and as the number of syllables of the English lyrics for paragraph level testing), **rhyme score (RS)**, which is defined as the average percentage of sentences within each paragraph that exhibit identical end

[3] https://music.163.com/

[4] https://github.com/ymcui/Chinese-LLaMA-Alpaca-2

[5] https://github.com/Sonata165/ControllableLyricTranslation

T	top-p	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
0.5	0.95	0.985	0.771	3.698	2.182	13.62	69.12
0.6	0.95	0.985	0.832	3.731	2.223	13.33	69.3
0.7	0.95	0.99	0.873	**3.76**	2.248	12.32	**69.43**
1	0.95	**1.0**	**0.901**	3.754	**2.325**	11.11	67.11
0.7	1	0.957	0.658	3.614	2.161	14.84	69.08

Table 3: Comparison of different sampling configurations (temperature and top-p probability).

Samples	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
1	0.862	0.385	3.061	2.074	12.79	67.94
80	**0.997**	0.862	**3.765**	**2.286**	12.32	68.84
40+40	0.99	**0.873**	3.76	2.248	12.32	**69.43**

Table 4: Comparison of no sampling, one-stage sampling, and our two-stage sampling strategy performance. 40+40 means the number of samples in two stages.

rhymes across all paragraphs, **basic and advanced translation quality** R_{bas} **and** R_{adv} as defined in Section 3.2, statistic machine translation metric **BLEU** (Papineni et al., 2002), and model-based machine translation metric **COMET**[6] (Rei et al., 2022). One caveat of BLEU is that it entirely depends on lexical form match and is sensitive to paraphrasing. On the other hand, COMET is robust and aligns much better with humans[7], so we mainly use COMET as the machine translation metric and report BLEU scores only for completeness.

4.2 Automatic Evaluations

The sentence-level performance of our generation models trained with several different recipes is reported in Table 1. In this experiment, we consider sentences in a paragraph as independent ones and set the desired length and rhyme according to our reference translation. We find that our dataset filtering strategy can largely improve translation quality by increasing all of R_{bas}, R_{adv}, and COMET. Also, after deleting the rhyme constraint in the prompt during inference time, generation results are still satisfactory even with slight improvements of R_{bas} and R_{adv}, though COMET slightly drops, partially due to the loss of length accuracy and therefore more misalignment with reference translation.

The whole-musical translation results are shown in Table 2, again indicating that our training strategy is effective and both our two training stages can boost performance. Comparing our final results and the baseline result, it is evident that we have achieved significant improvements across the majority of metrics. The only metric that ours is not as good as the baseline is the rhyme score since Ou et al. (2023) uses its so-called reversed decoding technique to benefit rhyme following at the cost of language quality, but our rhyme score is already high enough for most applications, especially considering that even English lyrics in a paragraph does not guarantee the same rhyme.

[6] the `Unbabel/wmt22-comet-da` variant

[7] COMET ranked 2nd in its alignment with humans among 20 metrics studied in Freitag et al. (2022), while BLEU only ranked 19th.

Samples	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
10+10	0.98	0.606	3.675	2.151	14.09	69.18
20+20	0.995	0.7	3.708	2.217	12.96	68.91
40+40	0.99	0.873	3.76	2.248	12.32	**69.43**
80+80	**1.0**	**0.906**	**3.777**	**2.269**	12.12	68.46

Table 5: Comparison of different numbers of samples in our framework, all using two sampling stages.

4.3 Ablation Study

In this section, we conduct ablation studies to show the effectiveness of our important design choices.

Model generation hyperparameters. Since we want to draw a large amount of samples for ensembling, the sampling configuration matters a lot. Table 3 shows the results if we change the temperature and top-p probability.

With a smaller temperature, the COMET score is generally better since outputs tend to have larger probabilities. However, the outputs are not diverse enough so the rhyme score is quite low. If we increase the temperature, the diversity becomes better at the cost of the COMET score. We believe the phenomenon that the COMET score hurts owing to a large temperature is especially significant in our constrained generation setting since acceptable solutions are often quite limited. We also ablate the effect of top-p sampling, finding that top-p sampling greatly improves sample diversity so both length accuracy and rhyme score improve, with a slightly better COMET score. We choose temperature $T = 0.7$ and top-$p = 0.95$ due to its best COMET score and high overall performance.

Our optimization framework. See Table 4 for the effectiveness of our optimization framework. If we do not use any optimization during inference and instead only sample once to get the final result, we will suffer from large drops across all metrics, especially in rhyme score. Compared with the simple one-stage strategy with equal computation resources (entirely using ensembling to fit a rhyme for a paragraph), adding the second stage can achieve a better rhyme score via more rhyme-conditioned samples.

Reward	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
no	**1.0**	**0.94**	2.972	2.064	13.8	67.08
yes	0.99	0.873	**3.76**	**2.248**	12.32	**69.43**

Table 6: The comparison of whether there are reward model terms in the inference loss function, signified by Reward in the heading row.

Model	Trained	LA	RS	R_{bas}	R_{adv}	BLEU	COMET
ours	no	0.844	0.574	3.731	2.159	12.49	68.1
ours	yes	**0.99**	**0.873**	3.76	2.248	12.32	69.43
Kimichat	no	0.944	0.669	**3.777**	**2.271**	15.98	**72**

Table 7: The comparison of closed-sourced Kimichat and both our untrained and trained model variants.

The number of samples in our framework. We can also freely adjust the number of samples. According to Table 5, roughly speaking, increasing the number of samples can improve rhyme score by a large margin. We find that 40 samples for both the first and second stages suffice for good performance while not being too time-consuming.

Reward model terms in the inference loss. We also show that reward models are helpful for the overall performance by contributing to the inference-time loss. With our best-performance configurations, deleting the reward model terms in the optimization process lets the COMET score drop by more than 2, as shown in Table 6. Compared with the one-sample setting in Table 5, here the no reward model setting has a slightly larger drop in COMET score since it attempts to fit rhyme score at the cost of translation results.

4.4 Additional Analyses

The scale of training data. Figure 4 shows that increasing the scale of training data can balance translation performance with length accuracy and rhyme score. Without training, the translation almost cannot follow the length and rhyme constraints. As we increase the size of the training set, length and rhyme accuracy keep increasing, at the cost of translation performance drops. This is reasonable because our training can help the model follow our constraints while the pre-trained knowledge can be diluted. Thus, we use 1.75M data to ensure high length and rhyme accuracy in the first training stage, and use high-quality filtered data to further improve translation quality in the second stage.

Using a larger closed-source LLM. Our method applies to any pre-trained LLM. It is then natural to explore the potential benefits of employing the state-of-the-art closed-source LLM for our purpose. We try Kimichat due to its good understanding and expression of Chinese and the result is shown in Table 7. The zero-shot translation quality of Kimichat is already better than our fine-tuned `Chinese-Alpaca-2-13B`, yet the length accuracy and rhyme score are not good enough. If we were able to apply our fine-tuning to the larger Kimichat, we would undoubtedly derive much better results, showing the performance scales well with the model size.

Method	Sentence-level				Paragraph-level	
	Fluency	Accuracy	Literacy	Alignment	Quality	Alignment
Ou et al. (2023)	2.88	2.53	2.37	2.48	2.08	2.92
Ours VER.1	3.09	2.6	2.45	2.69	2.31	2.75
Ours VER.2	3.25	2.64	2.54	2.6	2.27	**2.98**
Ours VER.3	**3.29**	**2.89**	**2.67**	**2.7**	**2.58**	2.96

Table 8: Human evaluation results. Our three versions correspond to those shown in Table 1, trained on different subsets: without filtering, with filtering, and with an additional second filtering.

Roughly decomposing our improvements. Compared with Ou et al. (2023), while achieving comparable performances of singability aspects, comprehensive experiments show that the success of our translation quality (approximated with the COMET score here) improvement mainly gives credit to two aspects. The first one is conducting dataset filtering using our trained reward models, which according to Table 1, 2, can contribute to about 2 points improvement of COMET score. Besides, the reward model terms in the loss function of our inference-time optimization framework give another 1.5 to 2 points improvement of COMET score, according to Table 5, 6.

4.5 Human Evaluation

We recruit 4 college students who are musical enthusiasts to do the human evaluation. We randomly sample 30 sentences and 12 paragraphs from our test set, let baseline and different versions of our model generate 120 sentences and 48 paragraphs, and ask another musical enthusiast to sing all generated results out. Subsequently, we let the evaluators assign scores on fluency, accuracy, literacy, and music-text alignment for sentence results, and overall translation quality and music-text alignment for paragraph results. We provide detailed scoring rubrics with examples and require the participants to adhere to our rules. This human evaluation can effectively demonstrate the subjective improvements of our methods over the baseline. See more details of human evaluation in Appendix D.

The results shown in Table 8 are generally consistent with our automatic evaluations. The im-

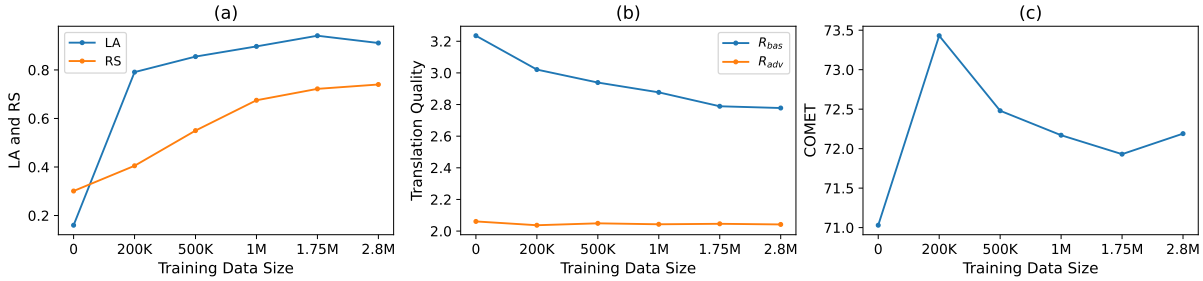

Figure 4: The changes of length accuracy, rhyme score, both basic and advanced translation quality, and COMET score if we change the training set scale.

Original lyrics	Ou et al. (2023)	Ours VER.3	Kimichat
You are sixteen going on seventeen Fellows will fall in line Eager young lads and rogues and cads Will offer you food and wine	你是十六个十七岁 伙伴们会结队 渴望年少顽童和部队 献给你餐酒一杯	你十六岁快要十七 兄弟们排成排 年少轻狂的无赖痞子 会为你提供美食	你十六岁，快十七了 同辈们排成行 热情小伙和恶棍流氓 将给你美食佳酿
Just because you find that life's not fair, it doesn't mean that you just have to grin and bear it! If you always take it on the chin and wear it Nothing will change.	只因你发现生活不公平 不代表只需要笑着忍痛 如果总是把它戴在你的头顶 不会变更	只因为你发现生活不公 不等于只能强颜而忍耐 如果总是硬着头皮强忍下来 永不更改	生活不公，何必默默承受 并不意味着你只能忍受 若你总是逆来顺受，默默承受 一切照旧

Table 9: Qualitative results for ours, baseline, and Kimichat. Translational errors and awkward phrases are underlined. Excellent lyrics are underwaved.

Original lyrics	Ours VER.1	Ours VER.2	Ours VER.3
Suddenly I'm flying company chaters	忽然间我飞去公司包机了	突然间我飞着公司的包机	突然间我正坐着包机飞往
Suddenly everything's high	突然什么都高涨	突然什么都高涨	突然一切都高涨
Suddenly there's nothing in between me and the sky	突然之间没有了我和天空相隔	突然之间隔着我和天空的天际	突然之间我和天空之间无屏障

Table 10: Qualitative results for our three versions corresponding to those shown in Table 1. They are trained on different subsets: without filtering, with filtering, and with an additional second filtering. Translational errors and awkward phrases are underlined. Excellent lyrics are underwaved.

provement of our VER.1 over the baseline and that of our VER.3 over the previous two versions demonstrate the effectiveness of our inference-time optimization and training dataset filtering.

Another thing worth mentioning is that although our rhyme accuracy is not as high as Ou et al. (2023), in human evaluation, our singability scores are all higher than the baseline, which means that our rhyming accuracy is already good enough for human listeners. People might pay more attention to how clearly we can hear the words in the lyrics given music, and that could be why we are seeing slightly improved results in text-music alignment.

4.6 Qualitative Results

In this section, we show a few representative qualitative results, with more results in Appendix C. For all Chinese translations, the translation errors and awkward phrases are underlined, and the excellent lyrics are underwaved.

Table 9 shows generation results of Ou et al. (2023), our model, and Kimichat with our optimization pipeline. All models achieved 100% length accuracy. For the baseline, though it has perfect rhyme, its translation quality is bad, with about one-third incorrect or awkward phrases. For our model and Kimichat, the rhyme accuracy reaches a satisfactory 75%, and has fluent, correct, and sometimes excellent translations. Table 10 demonstrates the effectiveness of our training recipe. With further finetuning with high-quality data, the percentage of awkward phrases is reduced, and more excellent translations emerge.

5 Conclusion

In conclusion, our work successfully balance translation quality and singability in musical translation. By leveraging trained reward models, a two-stage training approach, and an inference-time optimization framework, we have significantly improved upon previous methods. Our approach ensures that translated lyrics not only meet the criteria of fluency, accuracy, and literacy but also adhere to the critical constraints of length and rhyme. The substantial improvements over the baseline, as evidenced by both automatic and human evaluations, demonstrate the efficacy of our model in delivering high-quality translations that retain the essence of musical expression. This work paves the way for future advancements in the field, advancing cross-cultural appreciation of musicals.

Limitations

Although the current version of our reward models can already achieve good results, there is still room for further improvement by scaling the collected dataset and inviting more annotators to score sentence translations for less noise. By accessing more such resources, we believe the power of reward models could be even stronger, making the results more impressive.

Besides, we are translating at the sentence level due to the difficulty of tackling various constraints and composing sentences into a paragraph. Yet in some cases, neighboring sentence translations are not that compatible. Thus to further improve translation quality, we believe it is a promising direction to explore how to directly translate a paragraph.

Also, in this work, we only consider two of the most critical singability aspects for simplicity. In future works, it is possible to consider more fine-grained singability constraints to make our compositions more professional.

Ethics Statement

This work solves the task of musical translation, considering both translation quality and the singability constraints. Potential risks include not perfectly accurate translation results and thus could lead to misunderstanding if directly used in some scenarios.

The lyrics data are from the public Cloud Music[8] platform and are for research purpose only. The models we choose are only from public GitHub repositories with MIT licences. The dataset provided by Ou et al. (2023) is also used in its original intended way. For human evaluation, we collect evaluation scores without any personal information and we ensure the questionnaires do not contain any offensive statements.

[8] https://music.163.com/

References

Johanna Åkerström. 2010. Translating song lyrics : A study of the translation of the three musicals by benny andersson and björn ulvaeus.

B. Andersson, B. Ulvaeus, J. Craymer, and P. Dodd. 2008. *Mamma Mia! How Can I Resist You?: The Inside Story of Mamma Mia! and the Songs of Abba*. Orion Publishing Group, Limited.

Beatrice Carpi. 2020. A multimodal model of analysis for the translation of songs from stage musicals. *Meta*, 65(2):420–439.

Candice Jing Harn Chan. 2017. *The" visible" translator: challenges and limitations in musical translation*. Ph.D. thesis.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023. Improving translation faithfulness of large language models via augmenting instructions.

Hui Tung Cheng. 2013. *Singable Translating: A Viewer-oriented Approach to Cantonese Translation of Disney Animated Musicals*. Ph.D. thesis, Chinese University of Hong Kong.

Yiming Cui, Ziqing Yang, and Xin Yao. 2023. Efficient and effective text encoding for chinese llama and alpaca. *arXiv preprint arXiv:2304.08177*.

Lucile Desblache. 2018. Translation of music. *An encyclopedia of practical translation and interpreting*, pages 297–324.

L. Engel and H. Kissel. 2006. *Words with Music: Creating the Broadway Musical Libretto*. Applause Bks. Applause Theatre & Cinema Books.

Yuanhong Fei. 2014. 音乐剧翻译中的“信达雅”. 上海戏剧.

Johan Franzon. 2005. *Musical Comedy Translation: Fidelity and Format in the Scandinavian My Fair Lady*, pages 263 – 297. Brill, Leiden, The Netherlands.

Johan Franzon. 2008. Choices in song translation: Singability in print, subtitles and sung performance. *The Translator*, 14(2):373–399.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 46–68, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Fenfei Guo, Chen Zhang, Zhirui Zhang, Qixin He, Kejun Zhang, Jun Xie, and Jordan Boyd-Graber. 2022. Automatic song translation for tonal languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 729–743, Dublin, Ireland. Association for Computational Linguistics.

John Kenrick. 2010. *Musical theatre: a history*. Bloomsbury Publishing USA.

Chengxi Li, Kai Fan, Jiajun Bu, Boxing Chen, Zhongqiang Huang, and Zhi Yu. 2023a. Translate the beauty in songs: Jointly learning to align melody and translate lyrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 27–39, Singapore. Association for Computational Linguistics.

Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2023b. Eliciting the translation ability of large language models via multilingual finetuning with translation instructions.

Low. 2005. The pentathlon approach to translating songs. *Song and Significance*.

Peter Low. 2003. Singable translations of songs. *Perspectives*, 11(2):87–103.

OpenAI. 2023. Gpt-4 technical report.

Suezette Opperman, Marlie Van Rooyen, and Kobus Marais. 2018. An inter-semiotic approach to translation: Leonard cohen in afri-kaans. *Literator: Journal of Literary Criticism, Comparative Linguistics and Literary Studies*, 39(1):1–9.

Longshen Ou, Xichu Ma, Min-Yen Kan, and Ye Wang. 2023. Songs across borders: Singable and controllable neural lyric translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–467, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Karl Pearson. 1895. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I*, 58:240–242.

Olena Pidhrushna. 2021. Functional approach to songs in film translation: Challenges and compromises. In *SHS Web of Conferences*, volume 105. EDP Sciences.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Lucía Camardiel Sardiña. 2021. *The Translation of Disney Songs into Spanish: Differences Between the Peninsular Spanish and the Latin American Spanish Versions*. Ph.D. thesis, University of Hawai'i at Manoa.

Chen Si-yang. 2017. Practical strategies for devising singable song translations: A case study on wuhan university anthem translation. *Overseas English*.

S.L. Sorby, Hong Kong Baptist University. Department of English Language, and Literature. 2014. *Translating Western Musicals in Chinese: Texts, Networks, Consumers*. Hong Kong Baptist University.

Andrej Stopar. 2016. Mamma mia, a singable translation! *ELOPE: English Language Overseas Perspectives and Enquiries*, 13(1):141–159.

A. H. FOX Strangways. 1921. SONG-TRANSLATION. *Music and Letters*, II(3):211–224.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models.

Fan Xue. 2002. 歌曲翻译探索与实践.

Fu Yan. 1898. 译例言. 天演论译作.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2024. Tim: Teaching large language models to translate with comparison.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages.

A Dataset details

A.1 Dataset for Reward Model

We picked 11 musicals across various genres and spent about 20 hours extracting all the lyrics from their songs, breaking them down into paragraphs. Next, we used the Kimichat API to get initial translations for these paragraphs, tweaking our pipeline a bit: we kept the optimization but focused only on length and rhyme scores, as we did not have reward models yet. This gave us 15657 English lines. We then labeled 3938 of these lines in three different aspects, which took us another 30 hours. We divided the labeled data into training and test sets. Time and budget constraints meant we could not label everything, but what we did manage to label gave us pretty good results.

Our labeling metrics for human labeling is shown in Figure 5, 6, 7. We let human label in three aspects: fluency, translation accuracy, and literary. Each aspect has 4 levels of scores, and we give instructions and examples for each level to ensure consistency among human scores.

A.2 Test Dataset

We manually collect the lyrics from Cloud Music[9] and split them into paragraphs. The length constraint is obtained by counting the syllables of the English lyrics using the Syllapy library[10]. For testing BLEU and COMET scores, we collect the gold reference from human translations provided in Cloud Music. Our final test set consists of 409 paragraphs and 1741 lines.

B Implementation details

Generation Model Training Details. When training our `Chinese-Alpaca-2-13B` model, we use 1 epoch for both training stages. Training on 1.75M data samples takes about 9 hours using 8 80GB A100 GPUs. The codebase is adopted from the DPO GitHub repository[11] (Rafailov et al., 2023), which also supports supervised fine-tuning. We use the training batch size of 32 and keep all other hyper-parameters unchanged.

Reward Model Training Details. We also use `Chinese-Alpaca-2-13B` and the same codebase for reward models as for the generation model. For our basic translation quality reward model, there

[9] https://music.163.com/

[10] https://github.com/mholtzscher/syllapy

[11] https://github.com/eric-mitchell/direct-preference-optimization

are 471, 322, 971, and 2174 data samples with scores from 1 to 4. We upsample class 2 with a ratio of 1.5, downsample class 3 with a probability of 0.7, and downsample class 4 with a probability of 0.5. After adjusting the training dataset, we train our model with 5 epochs. Data downsampling means we keep each data sample with some probability, and data upsampling with a ratio p means we first keep one copy of the dataset and then conduct data downsampling with probability $p-1$ to derive additional data samples.

For our advanced translation quality reward model, there are 3104 samples with label 2 and 834 samples with label 3. We downsample class 2 with a probability of 0.4, upsample class 3 with a ratio of 1.5, and then train 5 epochs.

Inference-time loss function Details. We explain details in the inference-time loss function here

$$\mathcal{L}(y_1,\ldots,y_n)=\sum_i(\lambda_1[\mathrm{Rhy}(y_i)\neq\mathrm{Rhy}(y_n)]$$
$$+\lambda_2 D(\mathrm{gt}_i,|y_i|)-\lambda_3 R_{\mathrm{adv}}(y_i)-\lambda_4 R_{\mathrm{bas}}(y_i))$$

The length difference function is defined as

$$D(y,x)=\begin{cases}2(x-y) & \text{if } y\leq x,\\ y-x & \text{if } y>x.\end{cases},$$

and the four hyperparameters we use are

$$\lambda_1=2,\lambda_2=3,\lambda_3=1,\lambda_4=1.$$

We may change to other hyperparameters to gain slightly better results, yet in practice we find that this configuration can already achieve decent translation results.

C More results

Table 11 showcases the qualitative effect of using reward models in the optimization framework. without reward model terms, the translation quality significantly drops.

More cool results are shown in Table 12.

D Human Evaluation Details

We recruit 4 local college students who are musical enthusiasts from college's musical club wechat group. We randomly sample 30 sentences and 12 paragraphs from our test set, let baseline and three versions of our model generate 120 sentences and 48 paragraphs, and ask another musical enthusiast to sing all generated results out.

We let the evaluators assign scores on fluency, accuracy, literacy, and music-text alignment for sentence results, and overall translation quality and music-text alignment for paragraph results. We provide detailed scoring rubrics with examples and require the participants to adhere to our rules. The English version of instructions are shown in Figure 8, 9, 10.

It takes 3 hours for each annotator to do evaluations, and we pay them with a reasonable price for university students.

Original lyrics	Ours VER.3 with reward model	Ours VER.3 without reward model
Still strove, with his last ounce of courage, To reach the unreachable stars!	还在竭尽全力地奋斗 要飞到最远的星宿	拼了命的继续着奋搏 去到那不曾到过的
Well, let that lonely feeling wash away Maybe there's a reason to believe you'll be okay Cause when you don't feel strong enough to stand You can reach, reach out your hand	让那寂寞的感觉冲刷开 也许有一些理由相信你会好起来 当你感觉无力站起身来 你可以伸出手来	让那孤独感觉洗刷一洗 或许有理由相信你会过得很顺利 因为你太懦弱无法站立 伸出你的手可以

Table 11: The effects of using reward model terms in optimization pipeline. Translational errors and awkward phrases are underlined. Excellent lyrics are underwaved

Original lyrics	Generated Results
I am I, Don Quixote, the Lord of La Mancha My destiny calls and I go And the wild winds of fortune, will carry me onward Oh whithersoever they blow Whithersoever they blow, onward to glory I go	我是我，堂吉诃德，拉曼查的领主 宿命呼唤，我随之去 命运的狂风，将我带向未知旅途 任凭风儿吹向何处 任凭风向何处吹，我直奔荣耀而去
Hamilton faces an endless uphill climb He has something to prove He has nothing to lose Hamilton's pace is relentless He wastes no time What is it like in his shoes?	汉密尔顿面临无尽上坡路 他要证明什么 他无后顾之忧 汉密尔顿步履不停 他毫不耽搁 他脚下的路如何?
So let the sun come streaming in Cause you'll reach up and you'll rise again Lift your head and look around You will be found	就让阳光洒满房间 因为你会奋起再登攀 抬起头四处看看 必被发现
you will be popular! You're gonna be popular! I'll teach you the proper poise When you talk to boys Little ways to flirt and flounce	你会受到欢迎 你将会很有人气 姿势得体我来教 与男生谈笑 小动作挑逗撒娇

Table 12: More results generated by the third version of our model and Kimichat. Excellent lyrics are underwaved

Evaluation Criteria

Sentence Completeness

[Only look at the Chinese, not the English]

1. Content is absurd, illogical, or incomprehensible at a glance

> Thou art base and debauched as can be
> 你艺术基地就有多颓废
> To love, pure and chaste, from afar,
> 爱，纯且贞，远远地
> Timid and shy and scared are you
> 又胆怯害怕你是谁

2. Mostly complete sentences, but with hard flaws (**unacceptable**), such as the use of very inappropriate words, lack of necessary components, serious ambiguity, or disordered syntax

> Your life, little girl, is an empty page,
> 女你的生活是空的一页 （首字"女"很不合适）
> Cuz for the first time in forever
> 第一次长久以来的 （语序混乱，应为"长久以来的第一次"）
> And I know they'll take you home
> 我知道，带你回家 （缺少主语，"他们"带你回家）

3. Mostly complete sentences, no hard flaws (**acceptable**), but may have awkward wording or minor ambiguities, slightly off from normal Chinese sentences

> For fate to turn the light on
> 命运点亮希望光 （"希望光"用词略显尴尬）
> When you're broken on the ground
> 你在地上摔碎了 （"摔碎"用词尴尬）
> But his voice filled my spirit with a strange, sweet sound
> 但那声音注入我灵魂，奇妙甜美嗯 （结尾的"嗯"比较尴尬）
> In sleep he sang to me
> 他梦里对我唱（有歧义，在谁的梦里？）
> For my own sanity, I've got to close the door
> 为保心神平衡，我需关门远离 （说不清哪里不对，但怪怪的）

4. Very smooth, easily understandable

> Cause when you don't feel strong enough to stand
> 当你感觉站不稳的时候
> Even when the dark comes crashing through
> 就算那黑暗突然袭来

Figure 5: Metrics for human labeling, page 1/3.

> Your life, little girl, is an empty page,
> 姑娘你的生活，如空白纸张

Translation Accuracy

[Only look at the translation's fidelity to the original meaning, regardless of sentence completion, consider context]

1. More than 50% of the translation is incorrect, or a few key parts (such as active/passive voice, verbs) are translated incorrectly or missing, **unacceptable**

> Fellows will fall in line
> 兄弟长相厮守 （完全不对，应为"男人们会排队等待"）
> Tonight, we're gonna do ourselves justice,
> 今晚我们要做公正的自己 （关键部分不对，应为"今晚我们要为自己讨回公道"）
> I am sixteen going on seventeen
> 我是十六分继续十七分 （关键部分不对，应是"十六岁"、"十七岁"而非"十六分"、"十七分"）
> But now we're Ex-wives.
> 但现在，我们前妻。 （缺少谓语，我们"成为了"前妻）

2. Less than 50% of the translation is inaccurate, **barely acceptable** (allow for paraphrasing, allow for ignoring or changing a small amount of unimportant information)

> Don't know if I'm elated or gassy
> 不知我是欢喜还是气胀 （gassy在这里译为气胀不准确）
> And then I can go for a float
> 然后我能去漂浮了（"漂浮"不准确，应为游泳）

3. Basically accurate, but there is room for improvement, such as direct translation of English idioms without conveying the extended meaning, or adding a few small details would be better

> Where in the world have you been hiding?
> 你在地球上藏哪儿了？ （俗语，翻译成"你到底藏在哪儿了"就可以）
> What is it like in his shoes?
> 穿他鞋，感觉如何？ （俗语in sb's shoes，翻译成"如果我是他"更好）
> Sven, the pressure is all on you
> 史文，压力都在肩头 （小瑕疵，应当是"压力都在你肩头"）
> Couldn't keep it in, heaven knows I've tried
> 实在忍不住，竭力试过了（keep it in"忍不住"稍有点奇怪）

4. Very accurate in meaning (allow for paraphrasing, allow for ignoring or changing a small amount of unimportant information)

> I'll be dancing through the night
> 我会跳舞到夜晚
> But you're dying to try

Figure 6: Metrics for human labeling, page 2/3.

> 但是你想尝试

Lyric Quality

[Only look at the Chinese, don't need to consider sentence completion]

1. Not like real lyrics

> That one man, scorned and covered with scars,
> 那一个人被伤疤抹掉

2. Suitable to be used as lyrics, and has a certain literary quality

> it doesn't mean that you just have to grin and bear it!
> 并不表示你只需要笑着忍痛
> In dreams he came
> 梦中他来
> When you're broken on the ground
> 当你破碎在原地

3. Suitable to be used as lyrics, and has a certain literary quality

> For the first time in forever
> 因为好久没在生命里
> That one man, scorned and covered with scars,
> 那一人，受辱满身伤痕
> In dreams he came
> 梦中降临
> To run where the brave dare not go;
> 勇闯，无畏者所不至
> the ground is falling backwards
> 地面倒退飞逝

4. Very suitable to be used as lyrics, creative, expressive, and **eye-catching**

> To run where the brave dare not go;
> 跋涉，无人敢行的路
> My destiny calls and I go
> 这命运召唤我启航!
> The sweet caress of twilight
> 暮光轻抚，甜如诗

Figure 7: Metrics for human labeling, page 3/3.

Human Evaluation Instructions

Our project use large models for musical translation. Given English lyrics, the model will automatically generate corresponding Chinese translations. We have used different models and methods to generate some results, and we ask you to score these results according to our established rules.

The test is divided into two parts. The first part scores individual sentences on translation quality and singability respectively. This part consists of 120 questions. The second part scores paragraphs, requiring both consideration of the lyric text and its coordination with music. This part has 48 paragraphs. We provide reference audio for lyrics involving music coordination.

Part One: Single Sentence Scoring

You will receive: a line of English lyrics, a Chinese translation, a paragraph containing this English lyric; a raw song snippet, and a reference audio of the lyrics being sung.

What you need to do: First, based solely on the text, score on fluency, translation accuracy, and literacy; then listen to the original song snippet and the translated audio to score the coordination of the translated lyrics with the music. Scoring standards are as follows.

Fluency (Consider only whether the Chinese text is coherent and fluent)

- 1 point: Not human language - content is absurd, illogical, or incomprehensible at a glance

  > 爱，纯且贞，远远地

- 2 points: Partially coherent, but with serious flaws (unacceptable), such as inappropriate vocabulary, missing necessary components, serious ambiguity, or disordered syntax

  > 第一次长久以来的 (disordered syntax, should be "长久以来的第一次")

- 3 points: Mostly coherent, without serious flaws (barely acceptable), but with awkward wording or minor ambiguities, slightly different from normal Chinese sentences

  > 命运点亮希望光 ("希望光" is an awkward term)

- 4 points: Very fluent, easy to understand the meaning

  > 当你感觉站不稳的时候

Accuracy (Combine the paragraph to judge whether the lyric translation is accurate)

- 1 point: More than 50% of the translation is wrong, or a small number of key parts (such as passive voice, verbs) are translated incorrectly or omitted, unacceptable

  > Fellows will fall in line
  >
  > 兄弟长相厮守 (completely wrong, should be "男人们会排队等待")

- 2 points: Less than 50% of the translation is imprecise, barely acceptable (allowing paraphrase, allowing the omission or change of a small amount of unimportant information)

Figure 8: Instructions for human evaluation, page 1/3.

> Don't know if I'm elated or gassy
> 不知我是欢喜还是气胀 ("gassy" does not translate correctly here)

- 3 points: Basically accurate, but there is room for improvement, such as direct translation of English idioms without conveying the extended meaning, or could add some small details to improve

> What is it like in his shoes?
>
> 穿他鞋，感觉如何？ (The idiom "in sb's shoes" could be better translated as "如果我是他")

- 4 points: Very accurate in meaning (allowing paraphrase, allowing the omission or change of a small amount of unimportant information)

> To run where the brave dare not go
>
> 跋涉，无人敢行的路

Literacy (Consider only whether the Chinese text is suitable as a lyric)

- 1 point: Not like real lyrics

> 那一个人被伤疤抹掉

- 2 points: Can be used as lyrics, but plain and unremarkable, no highlights

> 并不表示你只需要笑着忍痛
> 当你破碎在原地

- 3 points: Suitable as lyrics, with a certain literary quality

> 因为好久没在生命里
> 那一人，受辱满身伤痕

- 4 points: Very suitable as lyrics, creative, expressive, and eye-catching

> 跋涉，无人敢行的路
> 这命运召唤我启航！

Single Sentence Evaluation of Lyric and Music Coordination

Mainly focus on three aspects:

- **Lyric word count**: Whether multiple words need to be crammed into one note, or one word corresponds to many notes? Generally, one note per word is the most suitable.
- **Pause**: Whether the pauses in the melody break up complete sentences/phrases? Ideally, the pauses in melody and semantics should coincide.
- **Misalign of tones and melody**: Is there a very serious reversal of words (hearing one word as another, such as "归来吧" heard as "鬼来吧")?

You don't need to consider translation accuracy here.

The audio examples for each score are in the file "Single Sentence Example.mp3".

Figure 9: Instructions for human evaluation, page 2/3.

- 1 point: The lyric word count is not perfect, it doesn't sound comfortable, there is room for improvement.

  > For the first time in forever
  >
  > 在人生中第一次 (incorrect length)

- 2 points: The lyric word count is very suitable, but the pause is very inappropriate or there is a very serious reversal of words.

  > is anybody waving back at me?
  >
  > 有没有人向我挥手回看 (There is a pause between "waving")

- 3 points: The lyric word count is very suitable, the pause is relatively suitable, the reversal of words is not very serious, but there are still strange-sounding places.

  > To right the unrightable wrong.
  >
  > 解决对不对的事情(“对不对”sounds strange, a bit of a reversal of words)

- 4 points: The lyric word count is very suitable, the pause is suitable, and the reversal of words is not serious.

  > For the first time in forever
  >
  > 永远的第一次体验 (the coordination of lyrics and music is good)

Part Two: Whole Section Scoring

You will receive: a section of English lyrics, a Chinese translation, and a reference audio of the translated lyrics being sung.

What you need to do: For the whole section, score the lyric quality and its singability.

Whole Section Comprehensive Evaluation

Lyrics Quality:

- 1 point: Most of the lyrics are not human speech, or most of the lyrics deviate from the original meaning.
- 2 points: Most of the lyrics are human speech, but there are still a few awkward places (unacceptable), such as inappropriate wording or translation errors.
- 3 points: The lyrics are barely acceptable, but there are still flaws.
- 4 points: It's hard to tell it's a translation, it seems like the original Chinese lyrics.

Text-Music Alignment:

- 1 point: Very poor coordination of lyrics and music, such as many sentences with incorrect word counts, very un-rhyming in rhyming sections...
- 2 points: The overall coordination of lyrics and music is acceptable, but there are some awkward problems, such as unreasonable pauses, serious reversal of words...
- 3 points: There are no major problems with the coordination of lyrics and music, but there are still flaws.
- 4 points: It's hard to tell it's a translation, it seems like the original Chinese song.

Figure 10: Instructions for human evaluation, page 3/3.